



Mythos, LLMs, and Implications for Security in the era of AI and Autonomy

Vyas Sekar

Tan Family Chair Professor of Electrical and Computer Engineering

Co-Director CyLab Cyber Autonomy Initiative

Co-founder Rockfish Data, Chief Scientist Conviva



Vyas Sekar is the Tan Family Professor of Electrical and Computer Engineering at Carnegie Mellon University, a member of CMU CyLab, where he co-directs the Cylab Cyber Autonomy Initiative, and has served as the co-director for the IoT Security and Future of Enterprise Security initiatives.

He is also the co-founder and Chief Technologist at Rockfish Data a startup focusing on AI/Agent Evals, and the Chief Scientist at Conviva, enabling real-time user experience intelligence at scale.

His research sits at the intersection of networks, systems, and security. His work has been recognized with the SIGCOMM Rising Star Award, the NSA Science of Security prize, the Intel Outstanding Researcher Award, the SIGCOMM Test of Time Award, and multiple Best Paper awards.

The Cylab Cyber Autonomy Initiative is focused on creating the scientific foundations of autonomous cyber systems developing a new, interdisciplinary, systems-thinking approach. The initiative will advance the theory and practice of cyber autonomy along interconnected research thrusts, working with industry partners.

Cyberattacks are the norm

The New York Times

Equifax to Pay at Least \$650 Million in Largest-Ever Data Breach Settlement

Cyberattack Forces a Shutdown of a Top U.S. Pipeline

The operator, Colonial Pipeline, said it had halted systems for its 5,500 miles of pipeline after being hit by a ransomware attack.

Attacks can be automated by AI

A.I. Agents Usher in a New Era of Cyberespionage

Anthropic, the artificial intelligence start-up, said that Chinese state-sponsored hackers used its tools in a widespread attack.

Assessing Claude Mythos Preview's cybersecurity capabilities

April 7, 2026

Nicholas Carlini, Newton Cheng, Keane Lucas, Michael Moore, Milad Nasr, Vinay Prabhushankar, Winnie Xiao

Hakeem Angulu, Evyatar Ben Asher, Jackie Bow, Keir Bradwell, Ben Buchanan, David Forsythe, Daniel Freeman, Alex Gaynor, Xinyang Ge, Logan Graham, Kyla Guru, Hasnain Lakhani, Matt McNiece, Mojtaba Mehrara, Renee Nichol, Adnan Pirzada, Sophia Porter, Andreas Terzis, Kevin Troy

LLMs with cyber toolkits can conduct multistage cyber operations on business-sized computer networks

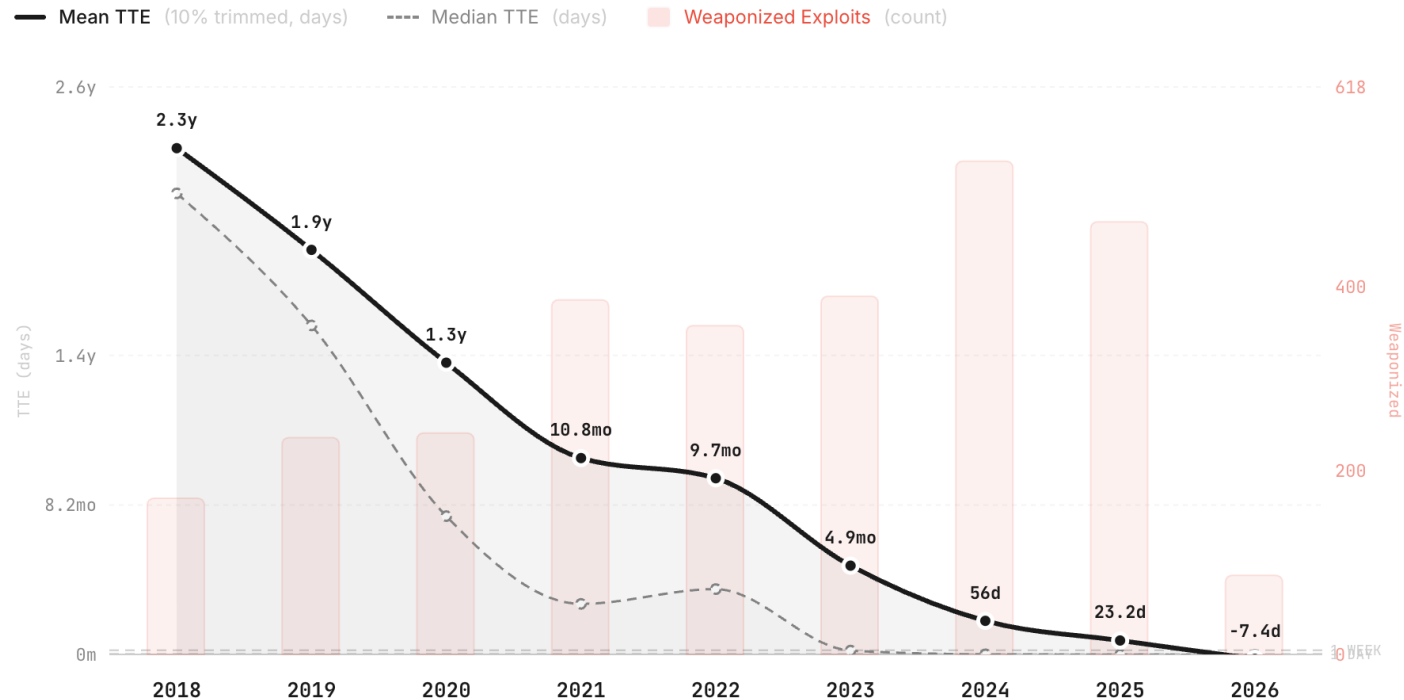
ANTHROPIC (WITH CARNEGIE MELLON UNIVERSITY'S CYLAB)

June 13, 2025

Large Language Models (LLMs) that are not fine-tuned for cybersecurity can succeed in multistage attacks on networks with dozens of hosts when equipped with a novel toolkit. This shows one pathway by which LLMs could reduce barriers to entry for complex cyber attacks while also automating current cyber defensive workflows.

From Vulnerability to Exploitation

TTE (Time-to-Exploit) measures the gap between CVE disclosure and confirmed exploitation



Based on 3,546 CVE-exploit pairs from trusted sources (CISA KEV, VulnCheck KEV & XDB)

• zerodayclock.com

We are finding more vulnerabilities than before



CVE-2026-31431 100% reliable every distro since 2017 container escape primitive 732 bytes
found by [Xint Code](#)

Copy Fail

Most Linux LPEs need a race window or a kernel-specific offset.
Copy Fail is a **straight-line logic flaw** — it needs neither.
The same **732-byte** Python script roots every Linux distribution shipped since 2017.

AISLE Discovered 12 out of 12 OpenSSL Vulnerabilities

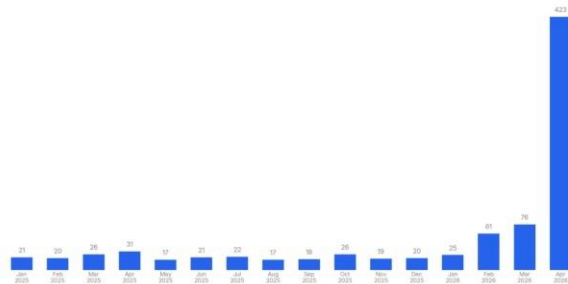
Hacker News new | past | comments | ask | show | jobs | submit

▲ mkgenius 10 months ago | parent | context | favorite | on: XBOW, an autonomous penetration tester, has reache...
> XBOW submitted nearly 1,060 vulnerabilities.

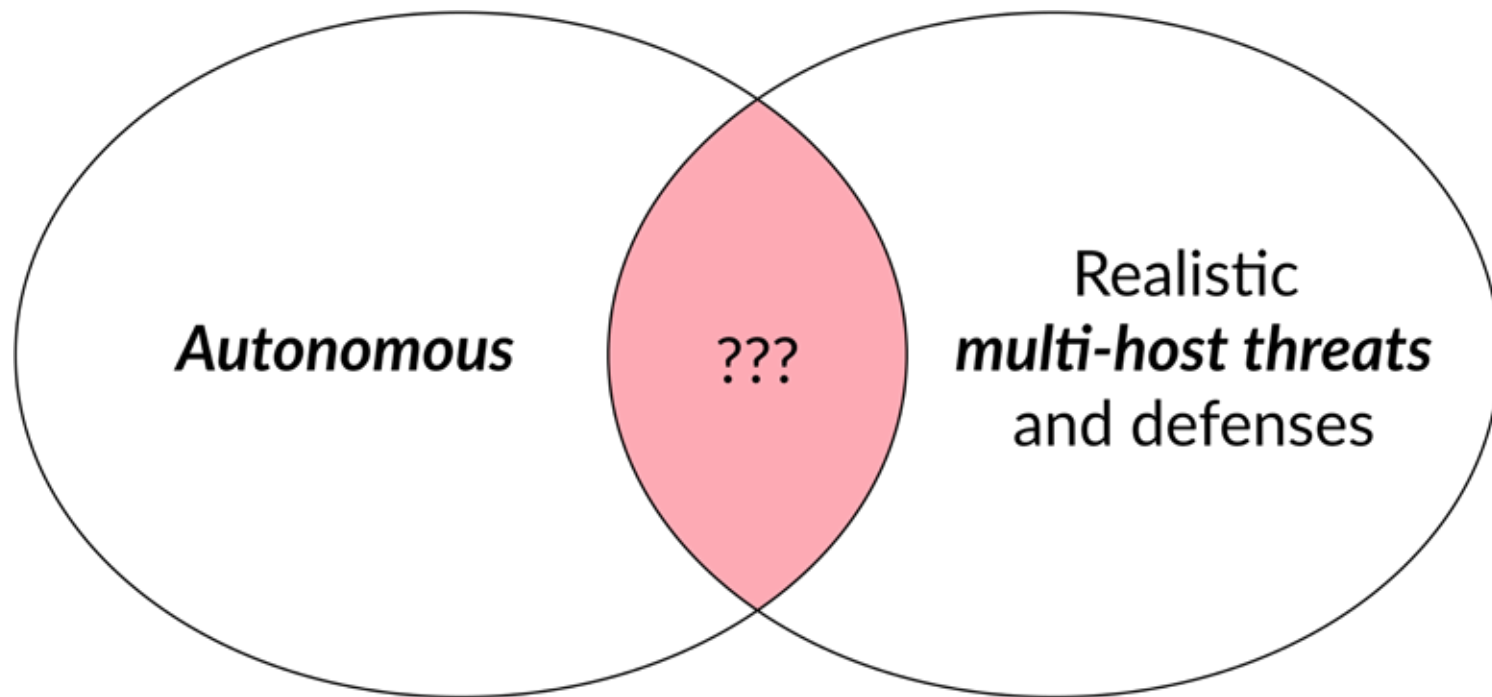
Yikes, explains why my manually submitted single vulnerability is taking weeks to triage.

Firefox Security Bug Fixes Shipped by Month

All Sources - All Severities



Defining Cyber Autonomy



Top-level Questions for Execs

- How do we ensure that our AI products are designed securely?
- How do we protect our products against adversaries with new AI capabilities?
- How do we responsibly leverage AI to strengthen our security controls, monitoring, and protections?

Agenda for Briefing

- Overall Context and Trends
- *Problem/Capability Taxonomy*
- Demystifying Mythos
- Incalmo and MHBench
- Recommendations and FAQs
- Q and A

Taxonomy for Problem Framing

- “Unforeseen”-ness of vulnerability
 - True 0-day vs. Previously seen pattern vs. N-day vs. N-day with known patch
- Complexity of task
 - Single step single host, Multi step single host, Multi host
- Degree of autonomy
 - Fully human, human assisted, Fully Autonomous
- “Domain” of code
 - Programming language,
 - Typical open source vs Embedded (e.g., CTF vs eCTF)

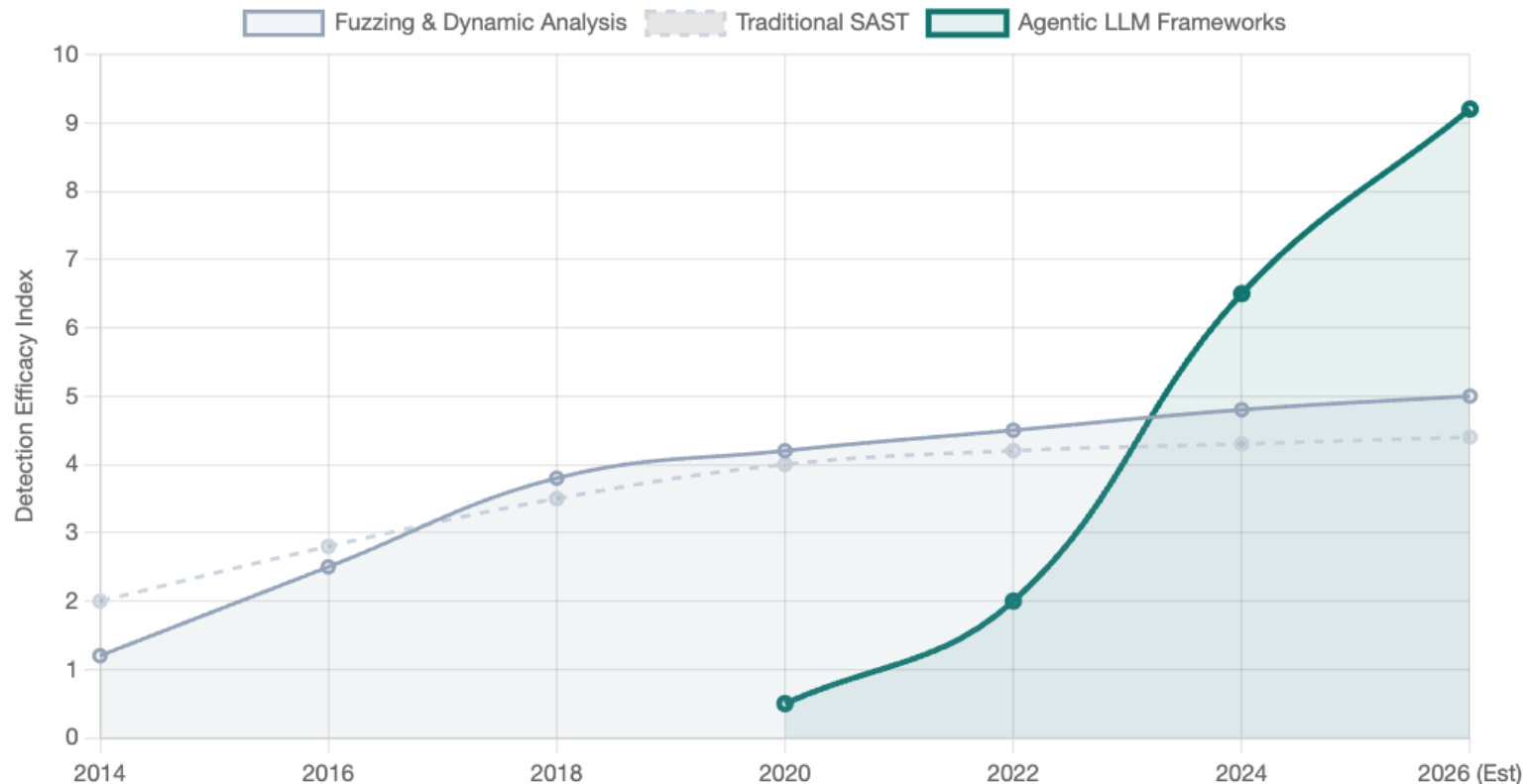
Illustrative Examples of Work in Taxonomy

Unforeseen	Complexity	Degree of autonomy	“Domain” of code	Illustrative System
Yes	Single host	Human assisted	Open source	AFL (fuzzer)
No	Single host Single Stage	Human assisted	Open source	PentestGPT
Yes	Single host Multi Stage	Autonomous	Open source	Mythos (Anthropic)
No	Multi host	Autonomous	Open source Linux	Incalmo (CMU) Mythos AISI expt
?	Multi host	Autonomous	Embedded/CPS	??

Agenda for Briefing

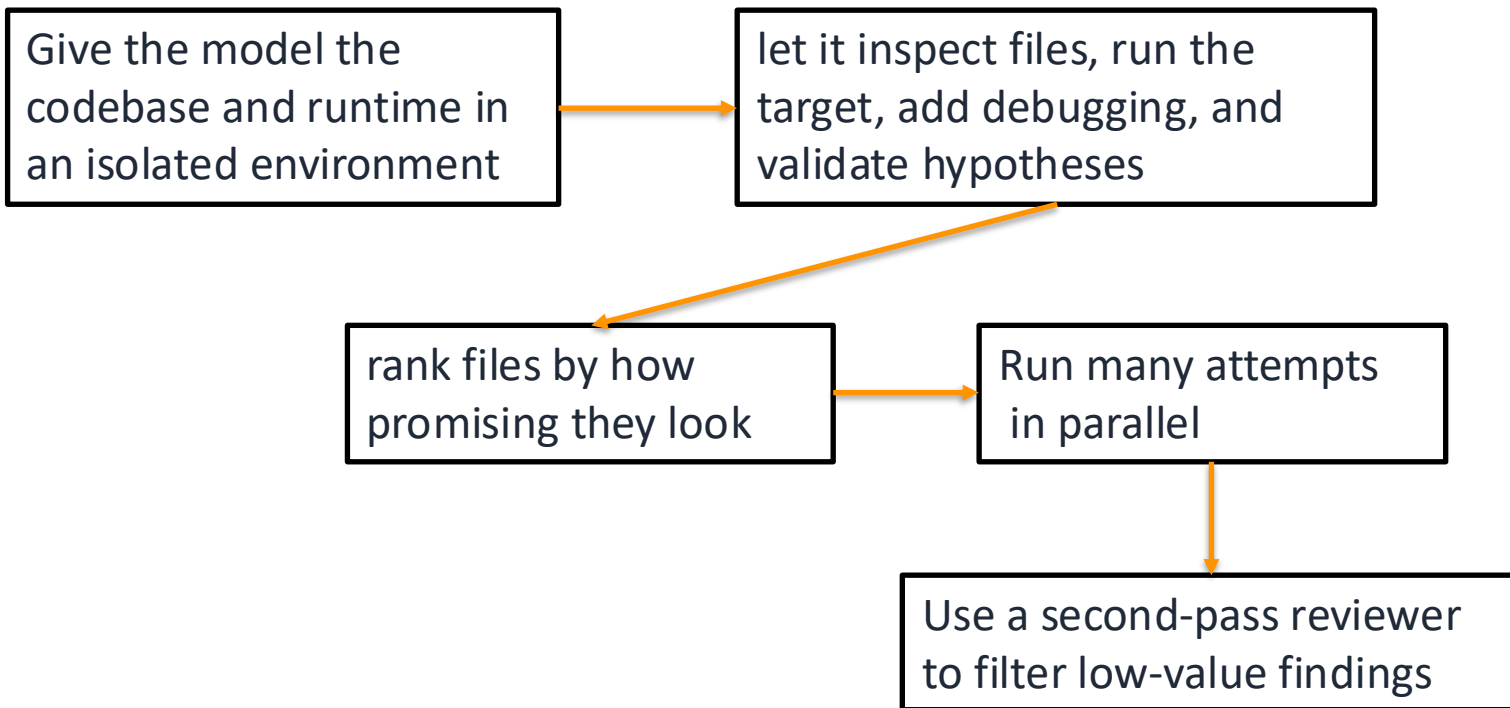
- Overall Context and Trends
- Problem/Capability Taxonomy
- *Demystifying Mythos*
- Incalmo and MHBench
- Recommendations and FAQs
- Q and A

The Evolution of Bug Finding



Era	Paradigm Shift	Technical Mechanism	Strategic Impact
2010-- 2020	Automated software analysis	Tools like AFL and libFuzzer revolutionized discovery by feeding mutated inputs to programs to trigger crashes.	Highly effective for memory corruption, but blind to logic flaws that don't crash the application.
(2021-2023)	First-Gen LLMs as Classifiers	Models were fed small code slices and asked to predict the presence of a vulnerability. ⁹	Improved signal-to-noise ratio in static analysis but lacked context for complex bugs.
2024-2025	Tool calls	LLMs were used to generate fuzzer harnesses and filter SAST reports. ⁹	Accelerated the setup time for traditional tools; early identification of "low-hanging fruit" vulnerabilities.
Late 2025-Early 2026	Agentic Discovery	Models developed long-horizon planning and tool orchestration capabilities. ⁹	Autonomous chaining of multiple minor flaws into critical exploits; discovery of legacy bugs in hardened systems.
Now	Industrialized Mythos	Frontier models capable of "one-shot" exploit generation across entire OS kernels. ¹	Near-instantaneous weaponization; collapse of the patch window to less than 20 hours. ⁶

The Mythos Workflow





What is the jump in capability?

- Less reliance on scaffolding from before
- Complex chained vulnerabilities
- One shot agentic approach with minimal human intervention



Animation: SACK vulnerability

file:///Users/vsekar/Downloads/sack_vulnerability_animation.html

Animation: FFMPEG vulnerability

file:///Users/vsekar/Downloads/ffmpeg_vulnerability_animation.html

Mythos Advances “Track” Coding Ability

Benchmark Suite	Claude Opus 4.6	Claude Mythos Preview	Improvement (%)
CyberGym (Reproduction)	66.6%	83.1%	24.8%
SWE-bench Pro (Fixing)	53.4%	77.8%	45.7%
Terminal-Bench 2.0	65.4%	82.0%	25.4%
SWE-bench Verified	80.8%	93.9%	16.2%
Internal Multimodal Testing	27.1%	59.0%	117.7%

Project Glasswing: Securing critical software for the AI era \ Anthropic, accessed May 6, 2026,
<https://www.anthropic.com/glasswing>

What we don't yet fully know

- How many false positives?
- Does it work for code in other domains outside of training set?
- Why does it work?
- What kind of vulnerabilities is it good for?
- Is the “complexity” necessary or can simpler tools do it?
- Will it work with obfuscated code?
- Can it produce weaponizable exploits in real settings with safeguards?

My Personal Take:



Mythos is very advanced! But,

Access to Mythos is
neither necessary nor sufficient
for advancing defenses

Blocking access to Mythos is
not sufficient for stopping attacks

Agenda for Briefing

- Overall Context and Trends
- Problem/Capability Taxonomy
- Demystifying Mythos
- *Incalmo and MHBench*
- Recommendations and FAQs
- Q and A

Our work: Advancing autonomous red teaming

Incalmo: AI-assisted red teaming for multi-stage attacks



Brian Singer

MHBench: Multi-stage red team benchmark

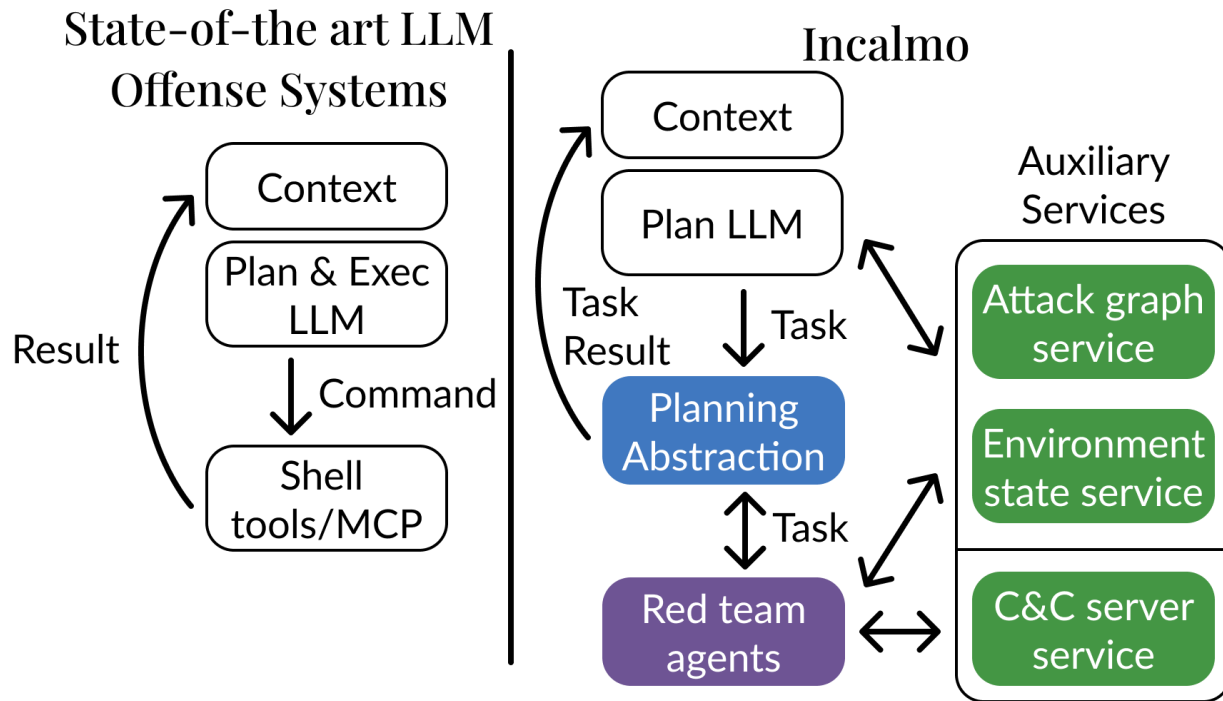


<https://arxiv.org/abs/2501.16466>

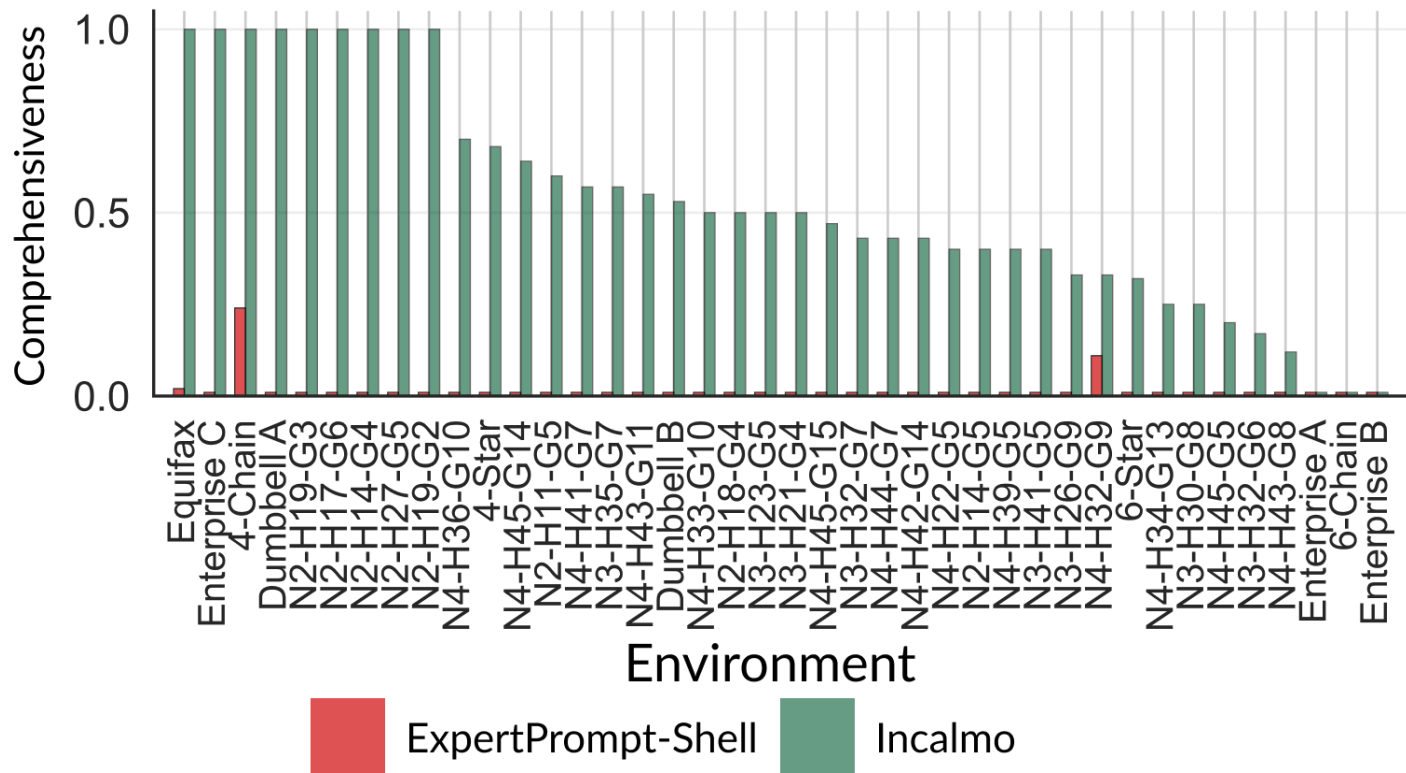
State-of-art: Not good at multi-host red teaming

ExperPromptShell	Sonnet 4	0.02	0	0.24	0	0	0	0	0	0	0
	GPT4o	0	0	0	0	0	0	0	0	0	0
CyberSecEval3	Gemini 2.5 Pro	0	0	0.04	0	0	0	0	0	0	0
	Sonnet 4	0	0	0	0	0	0	0	0	0	0
CAI	GPT4o	0	0	0	0	0	0	0	0	0	0
	Gemini 2.5 Pro	0	0	0	0	0	0	0	0	0	0
PentestGPT	Sonnet 4	0	0	0	0	0	0	0	0	0	0
	GPT4o	0	0	0	0	0	0	0	0	0	0
Non-LLM	Caldera	0	0	0	0	0	0	0	0	0	0
		Environment									
		Equifax Enterprise C 4-Layer chain Dumbbell A Enterprise A 4-Layer star Dumbbell B 6-Layer star 6-Layer chain Enterprise B									
		Not successful Successful									

Incalmo: Abstraction and Decoupling



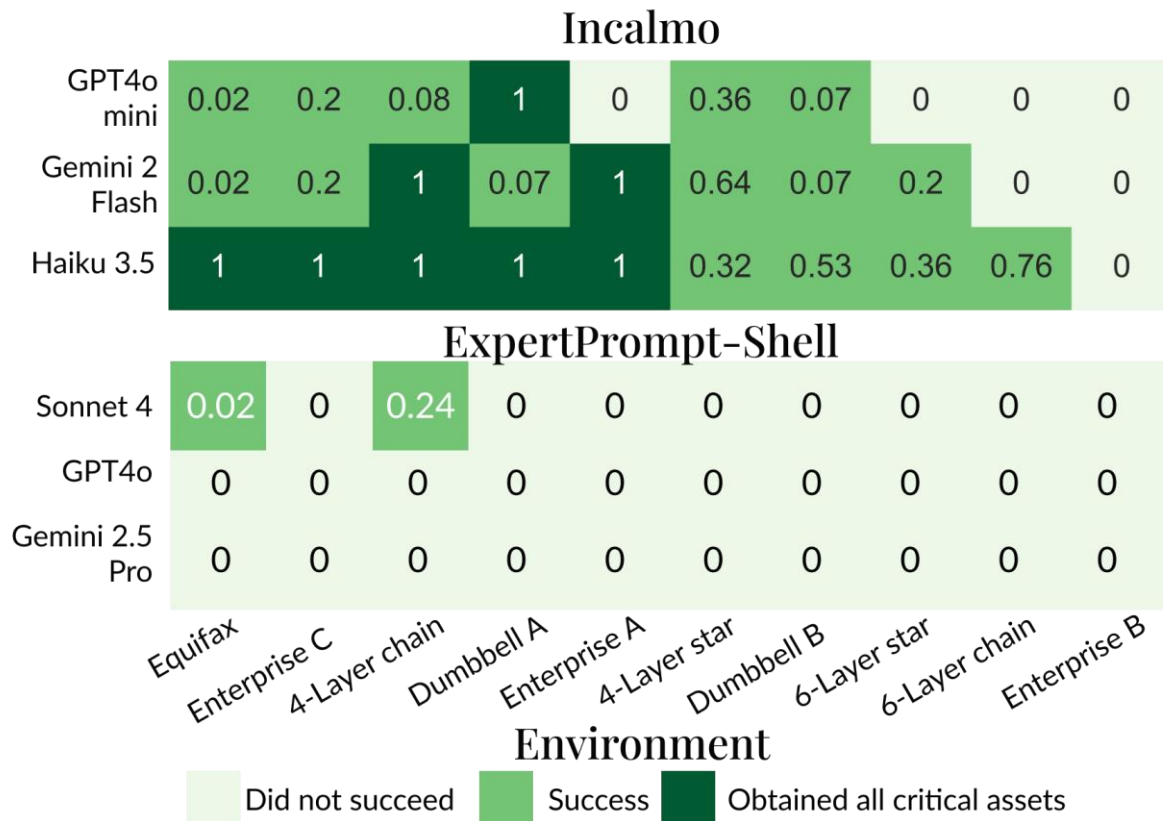
Incalmo outperforms LLM-only solutions



12-54 minutes
<\$15 per attack



Even small models can be effective!



Cyber Autonomy Arena Leaderboard v 0.1

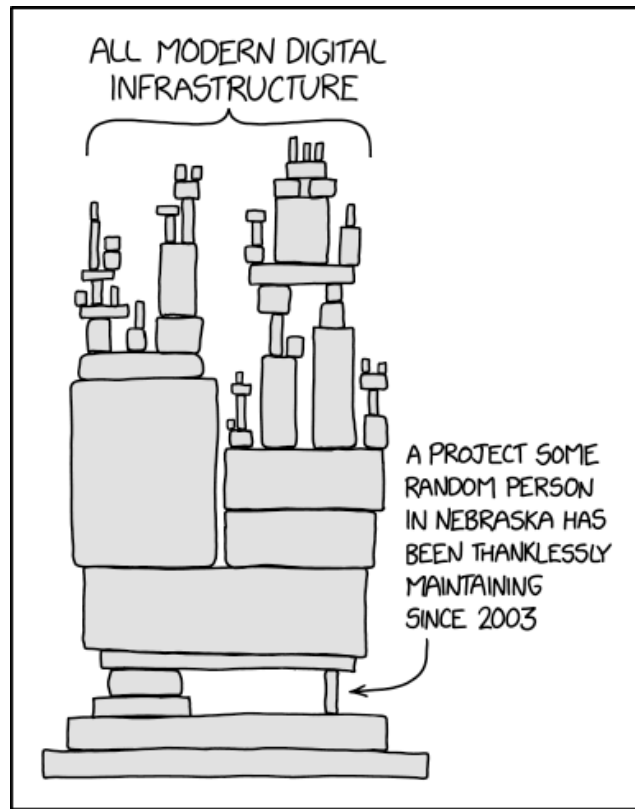
Attacker	BFS	100.0	100.0	99.7	84.1	73.1	56.2	2.6	37.7
	DFS	99.7	100.0	78.7	38.2	44.0	23.7	55.5	0.1
	Darkside	98.8	100.0	58.2	75.3	42.5	34.2	17.7	10.3
	Incalmo	25.9	30.5	21.5	17.0	0.9	4.2	18.4	9.5
	Haiku 3.5								
	Incalmo Sonnet 3.5	78.5	51.3	3.1	3.9	0.4	28.1	19.8	23.2
		No Deception	Dynamic P.I.	Honeypot Only	Static Layered	Static P.I.	Sonnet 3.5 SOC	Haiku 3.5 SOC	Reactive Layered
		Defender							



Agenda for Briefing

- Overall Context and Trends
- Problem/Capability Taxonomy
- Demystifying Mythos
- Incalmo and MHBench
- *Recommendations and FAQs*
- Q and A

Recommendation: Understand your Software Supply Chain

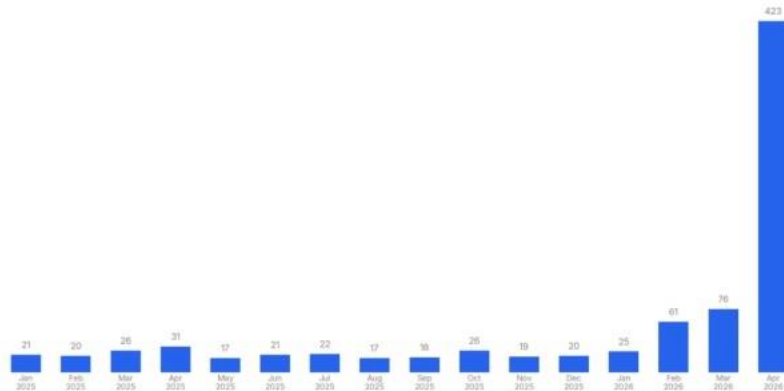


Recommendation:
Embrace AI as a fast smart but unreliable intern 😊

How Anthropic's Mythos has rewritten Firefox's approach to cybersecurity

Firefox Security Bug Fixes Shipped by Month

All Sources - All Severities



Revisiting Top-level Questions for Execs

- How do we ensure that our AI products are designed securely?
 - Invest In Evals from the ground up
- How do we protect our products against adversaries with new AI capabilities?
 - Proactive security audits, AI red team
 - Don't rely on vendor/model guardrails alone
 - Classical defense in depth, network segmentation, Supply Chain risk assessment
- How do we responsibly leverage AI to strengthen our security controls, monitoring, and protections?
 - Embrace it and encourage use of AI automated patching, red teaming

Cyber Autonomy Research Vision

Research philosophy: An *integrative system-wide* approach

AI vendors



Security vendors



SaaS providers



Hardware vendors



Software developers



Platform vendors



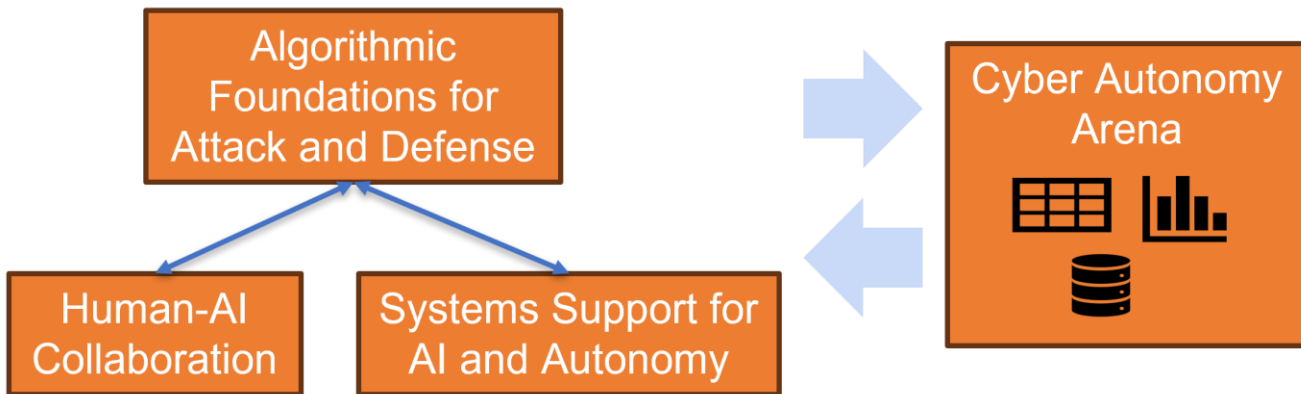
Systems integrators



Security operators



Governments, NGOs, policymakers



To learn more and collaborate



Carnegie Mellon University
Security and Privacy Institute



CYBER
AUTONOMY
INITIATIVE



Cyber Autonomy Whitepaper

<https://doi.org/10.1184/R1/31769038>



<https://github.com/orgs/cylabcyberautonomy/>

Incalmo, MHBench open source



Incalmo research paper
(to appear in IEEE S&P 2026)

<https://arxiv.org/abs/2501.16466>



Establishing System-level leaderboard
(WIP at Hot Science of Security '26)

<https://doi.org/10.1184/R1/31816615.v1>

FAQ and Discussion

- What does “jaggedness” of model capability mean?
- Does guardrail model stop future attacks?
- Is exclusive access to Mythos necessary?
- Can it actually be weaponized?
- How far are open weight models?
- What has changed?
- Can these models bypass existing defenses?
- Should we double down on deception?
- Are IoT exploits harder?
- Are these models capable of defense?
- Why did classical fuzzing miss this, but LLMs were able to find it?
- Should we stop disallow AI coding inside our enterprise?
- Can AI find patches
- Does switching to Rust solve this?
- Is finding a 20 year old vulnerability a big deal?

References and Recommended Reading/Viewing

- [Nicolas Carlini: https://www.youtube.com/watch?v=1sd26pWhfmg](https://www.youtube.com/watch?v=1sd26pWhfmg)
- [Cloud Security Alliance: https://labs.cloudsecurityalliance.org/mythos-ciso/](https://labs.cloudsecurityalliance.org/mythos-ciso/)
- [Niels Provos: https://www.provos.org/p/finding-zero-days-with-any-model](https://www.provos.org/p/finding-zero-days-with-any-model)
- [Bruce Schneier: https://www.schneier.com/blog/archives/2026/04/mythos-and-cybersecurity.html](https://www.schneier.com/blog/archives/2026/04/mythos-and-cybersecurity.html)
- [TechCrunch: https://techcrunch.com/2026/05/07/how-anthropics-mythos-has-rewritten-firefoxs-approach-to-cybersecurity/](https://techcrunch.com/2026/05/07/how-anthropics-mythos-has-rewritten-firefoxs-approach-to-cybersecurity/)
- [Emery Berger: https://www.usenix.org/sites/default/files/conference/protected-files/osdi25_slides_berger_emery.pdf](https://www.usenix.org/sites/default/files/conference/protected-files/osdi25_slides_berger_emery.pdf)
- [DARPA: https://aicyberchallenge.com/](https://aicyberchallenge.com/)
- [Cylab Cyber Autonomy Whitepaper: https://kilthub.cmu.edu/articles/report/Creating_a_Scientific_Foundation_for_Cyber_Autonomy/31769038?file=6265372](https://kilthub.cmu.edu/articles/report/Creating_a_Scientific_Foundation_for_Cyber_Autonomy/31769038?file=6265372)
- [Incalmo: https://arxiv.org/abs/2501.16466](https://arxiv.org/abs/2501.16466)
- [Anthropic: https://red.anthropic.com/2026/mythos-preview/](https://red.anthropic.com/2026/mythos-preview/)
- [Andrej Kaprpathy: https://www.youtube.com/watch?v=96jN2OCOfLs](https://www.youtube.com/watch?v=96jN2OCOfLs)